Article

Word Adjacency Graph Modeling: Separating Signal From Noise in Big Data

Western Journal of Nursing Research
1–20
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0193945916670363
wjn.sagepub.com
SAGE

Wendy R. Miller[1], Doyle Groves[2], Amelia Knopf[2], Julie L. Otte[2], and Ross D. Silverman[2]

Abstract

There is a need to develop methods to analyze Big Data to inform patient-centered interventions for better health outcomes. The purpose of this study was to develop and test a method to explore Big Data to describe salient health concerns of people with epilepsy. Specifically, we used Word Adjacency Graph modeling to explore a data set containing 1.9 billion anonymous text queries submitted to the ChaCha question and answer service to (a) detect clusters of epilepsy-related topics, and (b) visualize the range of epilepsy-related topics and their mutual proximity to uncover the breadth and depth of particular topics and groups of users. Applied to a large, complex data set, this method successfully identified clusters of epilepsy-related topics while allowing for separation of potentially non-relevant topics. The method can be used to identify patient-driven research questions from large social media data sets and results can inform the development of patient-centered interventions.

Keywords

epilepsy, Big Data, methods, informatics, machine learning

[1]Indiana University, Bloomington, USA
[2]Indiana University-Purdue University Indianapolis, USA

Corresponding Author:
Wendy R. Miller, Indiana University School of Nursing, 1033 East Third Street, Bloomington, IN 47405, USA.
Email: wrtruebl@iu.edu

This is the author's manuscript of the article published in final edited form as:
Miller, W. R., Groves, D., Knopf, A., Otte, J. L., & Silverman, R. D. (2017). Word adjacency graph modeling: Separating signal from noise in big data. Western journal of nursing research, 39(1), 166-185. https://doi.org/10.1177/0193945916670363

The current pervasiveness of social, web, and mobile application (app) use in people's lives has led to mass amounts of unorganized, user-generated data that are continually and organically updated by the public, and are often available for mining. The National Institutes of Health has defined these types of data as Big Data, and have created the Big Data to Knowledge (BD2K) initiative to support the development and testing of methods that can allow Big Data to influence health-related research (Margolis et al., 2014). By nature, Big Data sets are extremely large and heterogeneous, requiring processing applications that have not typically been used in health-related research. Yet, Big Data shows much promise in answering pressing health-related questions not possible with traditional research methods (Shaw, 2014). For example, electronic health records can allow researchers to draw conclusions about therapies and symptoms, length of stay, and other important outcomes using massive amounts of data, making results significantly more generalizable. There is thus an urgent need for the development and testing of methods to facilitate meaningful analysis of these data for health. With the right tools, Big Data could be leveraged to inspire or answer nursing research questions, particularly those related to self-management of chronic disease, which are optimized if patient-centered (Miller, Lasiter, Bartlett Ellis, & Buelow, 2015; Patient-Centered Outcomes Research Institute, 2016). Epilepsy, a chronic neurological disease affecting 2.9 million Americans, is one such disease that requires lifelong self-management (Epilepsy Foundation, 2016). Persons with epilepsy, like those with other chronic diseases, can benefit greatly from interventions developed based on organically generated data from patients with or affected by the disease.

Big Data and Chronic Disease

In the United States, 87% of adults use the Internet (Fox & Duggan, 2013). Moreover, 72% of Internet users report searching for health information online (Fox & Duggan, 2013; Wong, Harrison, Britt, & Henderson, 2014). Advances in mobile phone accessibility and technology have further facilitated the public's use of the Internet, social media, and mobile apps to gather health-related information, with 62% of smartphone users reporting that they use their phones to search for health information. Smartphones are highly accessible to people of all socioeconomic groups, levels of education, and ethnic groups (Smith, 2015).

The explosion of the use of the Internet, including social media and mobile apps, for health-related searches and discussions has created large, complex data sets that reflect the public's organic questions and concerns about health. These data are nationally representative, and are thus more

generalizable than the type of needs data typically used for designing nursing interventions. Yet, there remains a need to develop systematic methods that allow for the harnessing, organization, and meaningful analysis of these data to inform truly patient-centered programs, treatments, and other interventions that will improve human health. The National Institutes of Health's BD2K initiative was recently established to support the development of methods, software, and tools needed to analyze biomedical Big Data to improve human health. The National Institutes of Nursing Research is part of the BD2K initiative (Margolis et al., 2014). Nurse researchers are thus being called upon to engage in Big Data research, making it important for them to gain knowledge and experience in framing and answering research questions with these novel methods.

The potential power of social media in public health monitoring has recently been demonstrated, as existing computational and data visualization methods are being refined to manage the scope and volume of Big Data applied to health care. For example, Twitter has been used to track H1N1 influenza spread (Paul & Dredze, 2011; Signorini, Serge, & Polgreen, 2011), whereas Google searches have been correlated with dengue spread in tropical zones (Chan, Sahai, Conrad, & Brownstein, 2011). More recently, Twitter and Instagram data have been used to discover potential adverse- and drug-drug reactions (Correia, Li, & Rocha, 2016). This growing body of work further demonstrates the existence of health-related discussions in social media and other Internet-based platforms. However, there continues to be a lack of knowledge regarding the application of health-related findings from Big Data (Shaw, 2014). Nurse researchers have the opportunity to contribute to Big Data science by demonstrating ways in which these new methods can be applied to inform and help answer nursing research questions.

Morbidity and chronic disability from chronic disease accounts for half the health burden in the United States (Murray et al., 2013), and there is a critical need for the development of interventions that will positively affect the prevention and self-management of these diseases, with the ultimate goal of improving a variety of important outcomes—morbidity/mortality, quality of life, and health care resource utilization (Centers for Disease Control and Prevention, 2016). Chronic disease self-management is highly complex, and best described using a systems-level approach (Huang, Drewnowski, Kumanyika, & Glass, 2009). Disease management is affected by many variables such as individual characteristics, environment, and behaviors, and these variables cannot be fully captured or described with traditional types of data. Until now, most disease prevention and management interventions have been developed from research using these traditional methods, which are limited by access to relatively small samples that are often drawn from one or a

few geographical locations. Even well-powered multisite studies typically do not have sample sizes beyond hundreds of subjects, and do not afford researchers the opportunity to include input from millions of participants from around the world. Furthermore, traditional research methods involve some sort of investigator-initiated measurement(s), even in the form of qualitative interviews, while Big Data in the form of social media is a truer representation of patient voice and concern, particularly in anonymous platforms such as ChaCha (Priest et al., 2016).

Epilepsy is a chronic neurological disease that spans all age groups and affects one in 26 people in the United States. It is one of the 30 leading causes of disability in the United States (Murray et al., 2013) and is characterized by an enduring predisposition to generate epileptic seizures and by the neurobiological, cognitive, psychological, and social consequences of the condition (Fisher et al., 2005). Although surgery can be curative for some people with epilepsy, the majority of those affected by it are managed by pharmacological treatment. These various treatments often lead to undesirable side effects that affect individual tolerance leading to lack of adherence to medication and poor disease control, ineffective coping by individuals and families, and overall poor quality of life. Patients with epilepsy also state there is a stigma related to their disease and feel they do not have the ability to share concerns among friends and family. The challenges faced by people with epilepsy concerning disease management and stigma may be exacerbated by the structure and interpretation of interventional public policies (Burris, 2015), such as driving and accommodation laws, which attempt to balance broad public and workplace safety concerns against restrictions on individual rights and privileges. Therefore, the analysis of big data sets such as ChaCha, an anonymous question-answer service that has given rise to a big data set comprised of billions of queries, can reveal patient-centered concerns related to epilepsy in the patient-directed voice.

Purpose

The purpose of this study was to develop and test an analysis method to explore, using Big Data, salient health-related questions and concerns in people with chronic disease, using epilepsy as an exemplar. The specific aims were to implement Word Adjacency Graph (WAG) modeling on a question and answer ChaCha data set containing 1.9 billion anonymous queries to (a) detect clusters of epilepsy-related topics in a set of texts and (b) visualize the range of epilepsy-related topics and their mutual proximity to better understand the breadth and depth of particular topics and groups of users.

Method

ChaCha is a U.S.-based company that operates a free Short Message Service (SMS)–, mobile app–, and web-based question and answer service. The service is human-guided, in real time, and anonymous. Users of ChaCha are able to submit any type of question via the ChaCha platforms, and then receive a verified answer to that question. A user's questions are not shared with any other users, and each user can only see his or her own questions. The anonymity ChaCha users enjoy renders ChaCha data uniquely powerful in revealing individuals' authentic health concerns, thus providing researchers access to genuine patient voice. Although the public utilizes social media platforms such as Facebook, Twitter, and Instagram to discuss health and diseases, these platforms, although valuable, are typically not anonymous. In contrast, ChaCha provides a safe platform through which users can ask potentially stigmatizing or embarrassing health-related questions (Priest et al., 2016). Our research lab received a database of 1.9 billion ChaCha questions asked between January 2009 and November 2012. We received exclusive access to all raw data (questions and answers) generated on ChaCha during this time frame. We sought to apply a novel method, WAG modeling, to these data to reveal salient problems and needs in people with epilepsy.

WAG modeling is a type of social network analysis. Social network analysis is an approach designed to study social relations rather than individual attributes (Burt, 1978). WAG modeling involves viewing words in direct sequence from an input text. Words that are in direct sequence form pairs that in turn can form the nodes and edges of a network graph—nodes are words, and edges connecting those nodes are an indication that this word pair was found in the text. Over a full input corpus, or body of text, counts of edges show frequency of a word pair, and counts of nodes show frequency of individual words. Word pairs then chain together in clusters, and a visual layout of the resulting network graph often reveals several self-organized, self-describing subsets of words. Visual layout of the graph model was achieved using Gephi, an open source graph modeling tool (gephi.org). Partitioning was done first, and words were colored according to their resulting group number (modularity class number). The size of each words' representative circle was adjusted to reflect its betweenness centrality, or its "centralness" to the overall network structure. The ForceAtlas2 layout algorithm was used to arrange all words in a 2-dimensional space for visual presentation. The result is related words in the same group with the same color appearing close together, far from other less-related words and other groups.

Table 1. Epilepsy-Related Search Terms.

Whole Words and Phrases Searched in Fulltext		Stems Searched in Fulltext
spells	clobazam	epilep*
grand mal	phenytoin	seiz*
tonic clonic	carbamazepine	siez*
generalized seizure	gabapentin	
complex partial seizure	lacosamide	
temporal lobe seizure	oxcarbazepine	
anti-epileptic drugs	lamotrigine	
anti-epileptic drug	SUDEP	
AEDs	depression	
AED	anxiety	
seizure medications	suicide	
seizure medication	pregnancy	
seizure meds	pregnant	
levetiracetam		

Note. Drugs listed in table were also searched using trade names.

Partitioning techniques (e.g., Louvain modularity) can further be applied to the network graph to compute these precise word groups on a large scale without necessarily requiring the visual layout step. Given a graph model, partitioning is a process by which the graph is divided into subsections, where each subsection has many connections among its members, and few connections to members in other subsections, revealing the natural clusters present in the graph. In this application, those clusters of similar words equate to individual subtopics within the full text. We used the modern, robust Louvain modularity partitioning algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008).

Procedures

Our institutional review board determined that our study met exempt status given that it involved retrospective analysis of de-identified data. From the initial 1.9 billion ChaCha questions, an epilepsy-related subset was identified by searching for questions containing epilepsy- and seizure-related terms (see Table 1). Approximately 300,000 included strongly related terms, and another 12 million contained health terms more weakly related to epilepsy. The stronger-signaled set of 300,000 questions were fully modeled first. Each question was split into token words (an individual word in a sentence, often separated by spaces, tabs, and punctuation from other tokens) and stop words

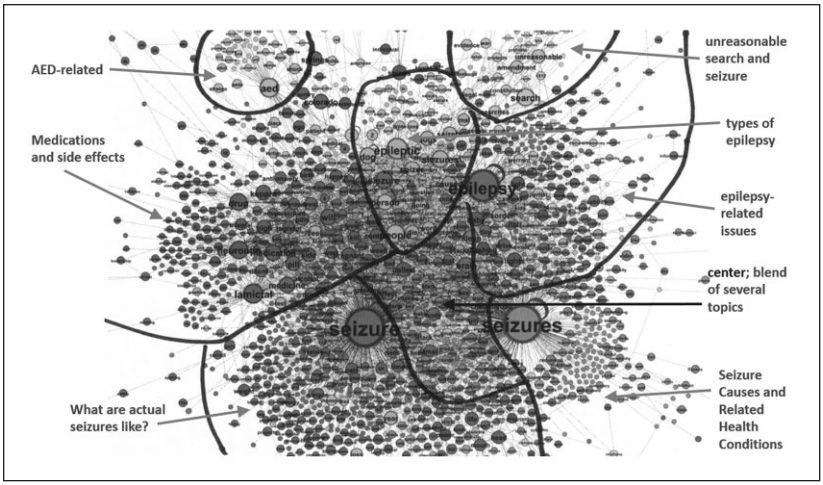

Figure 1. Initial WAG model from 300,000 questions.
Note. WAG = Word Adjacency Graph.

(words of very little informative value on their own; things such as a, the, and that string other high-meaning words together). Remaining words were considered in sequence to build a summary table of words, 2 grams, and 3 grams (n-gram is a string of words of "n" length; a 1 gram is a single word, and 2 gram is a two-word phrase, for example). Two grams occurring at least 10 times were used to build a network graph model (see Figure 1). We chose a minimum of 10 occurrences to find a balance in two respects. First, to choose a number that represents enough signal, 10 times is better than 1 or 2 times to feel that the phrase occurs sufficiently frequently in the wild. Second, a higher minimum occurrence limit results in fewer words that can be included in the model. A minimum of 10 mentions yielded 1,582 words and 4,429 word pair connections, sufficient in both the high quantity of words for observing richly detailed groups, and the high quantity of mentions for each.

The graph model was refined by performing partitioning, then reviewing resulting subsections of the graph manually for appropriateness. A few large topics were quickly observed to be irrelevant, such as "seized" when discussing war and territory, rather than seizure, and "spells" when discussing magic, rather than spells related to epilepsy (Figure 2). Several very small groups of two or three words were also discarded, as they represent content unrelated to epilepsy, and were completely unconnected to the single, central mass of 15 detailed topics.

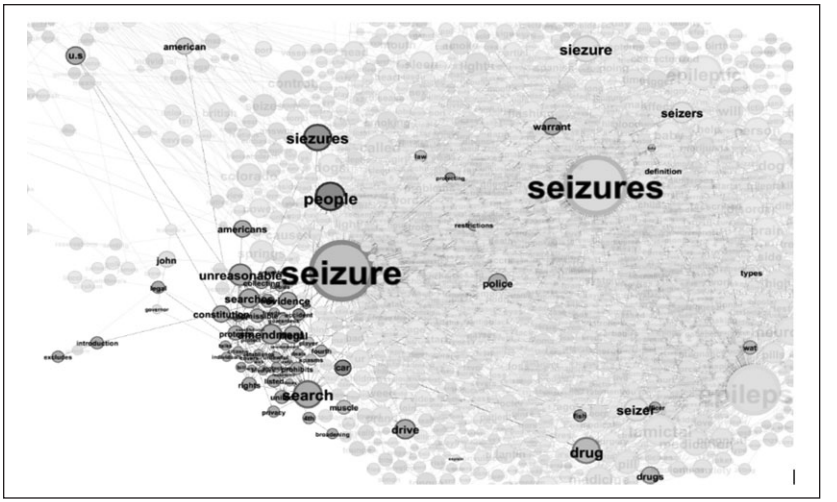

Figure 2. Close-up view of relevant cluster.

The remaining 15 topics were reviewed manually also, to create the short descriptions of each in Table 2. The authors' collective opinion of these groups further designated 11 as related to epilepsy, and four much less relevant. Highlighting one cluster shows its self-description, and words at its edge hint at topics it shares with adjacent topic neighbors (see Figure 3). Very early in the process, clusters (a set of words that are closely related) of unrelated content were observed and earmarked for removal on subsequent layout procedures (see Figure 3).

Results

The above-described process rendered 15 individual clusters, 11 of which we consider relevant to the core topic of epilepsy, and four that are much less relevant. Thus, the initial WAG modeling applied to this large data set has allowed successful separation of relevant and non-relevant epilepsy questions (see Table 2). Table 3 represents labeled clusters with supporting user queries.

Clusters 1 and 2 appear to represent general questions about seizures, including their definition and description. The third cluster includes questions that relate to named anti-epileptic drugs, as well as their side effects. An example includes a user asking if mood disturbances can be caused by the medication levetiracetam (an anti-epileptic drug). Clusters 4 and 5 include questions related to the definition of epilepsy and epileptic, as well as life

Table 2. Manually Assigned Cluster Labels.

Cluster Number	Description
1	"seizure"; what actual seizures are like
2	"seizures"; causes and related health conditions
3	medications by name, and their side effects
4	"epileptic" (both people and dogs); types of epilepsy, seizures and fits
5	"epilepsy"; peripheral life issues—driving, drinking, marines, etc.
6	AED-related[a]
7	effects on the brain and body
8	flashing lights, and some history questions
9	pregnancy-related
10	unreasonable search and seizure[a]
11	anxiety, heart attacks, moods (and some pokemon, which is actually relevant)
12	marijuana-related
13	sex, time, energy drinks
14	games, battles, wars[a]
15	pills, movies[a]

a. Indicates potentially non-relevant cluster.

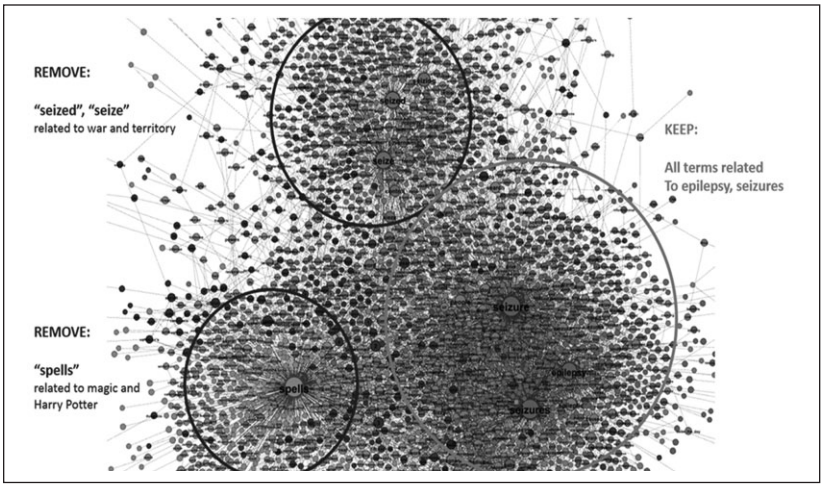

Figure 3. Close-up view of irrelevant versus relevant clusters.

Table 3. Cluster Descriptions and Supporting User Queries.

Cluster(s)	Description	Example Queries
1 and 2	General questions about seizures, including definitions	What is epilepsy? What is a seizure? Can you have a seizure but not shake? What's a complex partial seizure?
3	AEDs and side effects	Can levetiracetam make you nervous? Why do seizure pills make you depressed? Can phenytoin make you really tired?
4 and 5	Definition of epilepsy and life issues	How do you know if you qualify as having epilepsy? Does one seizure mean epilepsy? Does epilepsy run in the family? How long till I can drive with seizures? Can you be fired because of seizures? Is alcohol okay with epilepsy?
7	Effect of epilepsy on the body	Can you bite your tongue off during a seizure? Can you get a bad head injury if you have a seizure?
8	Seizure triggers in the form of strobe/flashing lights/ games	Can flashing lights give a person with epilepsy a seizure? If I have seizures and play Pokemon™ am I going to have a seizure?
9	Epilepsy/seizures and pregnancy	When you are pregnant does epilepsy get worse? Can pregnancy make you have a seizure?
11	Comorbidities	Why do epileptics get depressed? Can epilepsy make your heart stop? Why do seizures make me feel nervous?
12	Marijuana as treatment/ induction for epilepsy	Can pot make your seizures worse? Is it true that marijuana can stop seizures?
13	Epilepsy and sexual health/ triggers	Can having sex give a person with epilepsy a seizure? Can a person with seizures have sex? Can I drink energy drinks if I have epilepsy?

changes related to epilepsy (e.g., driving, employment, ability to join the military). Users were specifically interested in whether epilepsy is hereditary, the differences between different types of epilepsy (such as juvenile myoclonic epilepsy and generalized epilepsy), and driving, lifestyle (e.g., drinking alcohol), and employment-related restrictions associated with epilepsy. The sixth cluster of questions contains the acronym "AED," which stands for anti-epileptic drug. However, nearly all of the queries in this cluster appear to relate to an automatic external defibrillator, which also goes by the acronym AED, and thus this cluster is not relevant to the current study.

Cluster 7 includes questions relating to the effect of epilepsy and seizures on the body, particularly the brain and tongue. Cluster 8 includes a mix of questions related to seizure triggers (flashing and strobe lights), as well as what appear to be history questions regarding the seizing of nations by governments or historical figures, most notably Napoleon. The ninth cluster contains queries related to epilepsy, seizures, and pregnancy, though questions in this cluster were not related to anti-epileptic drug effects on pregnancy (these questions were contained in Cluster 3). Cluster 10 is irrelevant to epilepsy, containing questions related to search and seizure, as by police or military personnel. Cluster 11 questions pertain to comorbidities of epilepsy—anxiety, depression, heart problems, and, interestingly, Pokemon, which is a video game. All questions in Cluster 12 are marijuana-related and relate the drug's ability to stop or induce seizures and to treat epilepsy. Cluster 13 includes a mix of questions related to sex and epilepsy, energy drinks and seizures, and time—often referring to the duration of seizures. Clusters 14 and 15 are both irrelevant to epilepsy, as they relate to games, battles, and wars (14), and movies and non-epilepsy-related pills (15).

Next steps involved submitting the larger source material (12 million queries) to analysis to uncover additional peripheral topics beyond these core 11 from the smaller, strong-signal 300,000 question source set. For the larger data set of 12 million questions, we chose a new minimum of 100 occurrences, resulting in a model with 5,342 words and 29,312 word pairs. This set was more than 3 times as many words as the earlier, smaller model, but with a lower quality partitioning result—modularity value of 0.36 versus 0.48 earlier. It also yielded 10 significant groups, fewer than the 15 from the smaller model, so a smaller range of discovered topics. We expected, from this analysis of the larger data set, that we would find words related to epilepsy that were also related to other topics (health issues, lifestyle, and behaviors) and words related to epilepsy, although not directly adjacent to words around other topics, that may occupy adjacent content space that is relevant. However, no obvious examples of either of these were found on our initial attempt at a more broad ("zoomed out") analysis of the data using the WAG modeling method.

Discussion

We achieved our purpose of testing a method for identifying salient health concerns within a broad-scope Big Data set. We were able to refine and apply WAG modeling to answer health-related questions in a Big Data set. Applied to a large, complex ChaCha data set, WAG modeling was successful in identifying clusters of epilepsy-related topics within the data, while also allowing for the separation of potentially non-relevant topics. Thus, this method has resulted in an initial, organized, somewhat broad list of epilepsy-related areas of concern to users.

The extant literature demonstrates that the public, and even people with epilepsy, have a low level of knowledge and understanding of epilepsy and seizures (Hesdorffer et al., 2013), which is reflected in Clusters 1, 2, and 4 in this analysis. Coupled with extant literature regarding knowledge deficits in persons with epilepsy (Miller, 2014; Unger & Buelow, 2009), clusters emerging from this analysis suggest that, at least as of 2012, there remains a need to implement a widespread and effective public awareness campaign regarding epilepsy, and for persons diagnosed with epilepsy (and their families) to receive detailed information about the epilepsy diagnosis. This dearth of epilepsy information and education has been noted in prior qualitative studies (Miller, Bakas, & Buelow, 2013; Unger & Buelow, 2009). Relatedly, Cluster 11 questions were mostly about comorbidities associated with epilepsy, such as other physical (heart problems) and mental (anxiety, depression) illnesses that can coincide with epilepsy. Many mental and some physical illnesses are highly associated with epilepsy (Epilepsy Foundation, 2016), and thus persons with epilepsy or their loved ones may not be receiving this very well-known and important information from care providers.

Anti-epileptic drugs very commonly cause adverse side effects, and these side effects are a chief reason for medication non-adherence (Epilepsy Foundation, 2016). The frequency with which users were asked about side effects of these medications (≥10,000 queries) in Cluster 3 indicates that people taking these medications, or their loved ones, may be unaware or uncertain of typical side effects. In addition, questions included in this cluster have the potential to elucidate side effects and drug interactions not currently known. For example, 29 unique users asked if the medication lamotrigine causes weight gain. According to manufacturer information regarding lamotrigine, weight gain is not a documented side effect of this medication (Epilepsy Foundation, 2016). It must thus be pondered whether weight gain is a side effect of this medication that was not found in the initial studies leading to its approval for use. This particular finding reveals the potential of social media data to inform knowledge regarding drug-drug interactions,

which has been demonstrated in Twitter and Instagram (Correia et al., 2016), as well as medication side effects, and has specific implications for nursing practice. Nurse researchers working with advanced practice and direct care nurses should consider social media data analysis as a viable method of validating yet unknown drug-drug interactions and side effects based on patient reports in the clinical or community environment. Furthermore, more than 300 users posed questions about birth control and anti-epileptic drugs. Many anti-epileptic drugs decrease the effectiveness of birth control (Carl, Weaver, & Edgerton, 2008), and the frequency of questions about this topic suggests that women taking birth control and anti-epileptic drugs (or the partners of those women) may be unsure about the effect of seizure medications on birth control, despite having received both medications from health care providers. Cluster 9 questions, which related to the effect of epilepsy on a baby and mother during pregnancy, further indicate the salience of women's health issues in epilepsy.

Questions in Cluster 5 include those related to common life issues for people with epilepsy—driving, employment, alcohol consumption, and ability to enroll in the military. All of these questions are supported by quality of life studies in people with epilepsy (Hesdorffer et al., 2013; Miller et al., 2013; Unger & Buelow, 2009). Interestingly, an existing intervention aimed at improving the quality of life of people with epilepsy through stress reduction, medication adherence, and improved sleep quality has not demonstrated statistically significant effects on these quality of life issues (DiIorio, Bamps, Walker, & Escoffery, 2011). The presence of these issues in the data set indicates the continued need to address the drastic life changes associated with epilepsy, and to reconsider the ways in which epilepsy providers, researchers, and organizations currently address these issues. As well, our results suggest the need to further investigate the chief needs and concerns of people living with epilepsy.

Questions in Cluster 7 reflect users' questions about the effect of epilepsy on the body, as well as injuries to the body that may cause epilepsy. For example, there were nearly 500 questions related to a person swallowing his or her tongue during a seizure. Although it has long been known that it is impossible for a person to swallow his or her tongue during a seizure, this myth persists (Epilepsy Foundation, 2016), and these tongue-related questions in this data set demonstrate this continued lack of knowledge regarding seizure first aid. Myths surrounding tongue swallowing during seizures have, in the past, led to seizure first aid actions that are dangerous to people with epilepsy. Inserting a spoon or other object into a person's mouth during a seizure can cause significant injury (Epilepsy Foundation, 2016).

More than 500 questions related to seizure triggers in the form of strobe or flashing lights in Cluster 8 demonstrate users' interest and potential lack of knowledge about this subject. For example, the majority of these queries related to whether flashing or strobe lights could cause seizures. Flashing and strobe lights are well-documented as seizure triggers, again suggesting the need to reexamine the ways in which basic information about epilepsy is being disseminated to people with epilepsy and to the public. Other trigger-related questions were seen in Cluster 8, in the form of 10 unique questions about a specific Pokemon game causing seizures. There is no documentation of Pokemon video games being especially likely to trigger seizures, though a specific Pokemon game was asked about in terms of its ability to cause seizures by 10 different users. This finding is interesting in a way similar to that of users inquiring about side effects to medications that are not documented in the scientific literature—is there a yet unknown aspect of this particular game that is seizure triggering? This finding also points to an interesting potential for this approach as an early folklore detector to aid in preventing a viral outbreak of misinformation. Finally, what appear to be trigger-related questions regarding energy drinks and alcohol were included in Cluster 13, with users asking if it is safe to drink these beverages and if they can precipitate seizures. Caffeine and alcohol are common seizure triggers (Epilepsy Foundation, 2016), again suggesting the lack of knowledge of basic epilepsy self-management information by these users.

That an entire cluster of marijuana-related questions emerged from the data set was not surprising. Marijuana has been associated with epilepsy in both its ability to provoke seizures and, in the last decade, the potential ability of its derivatives to prevent them. Cannabis is currently not approved for treatment of epilepsy, though there are ongoing studies testing its effectiveness in treating the condition (Epilepsy Foundation, 2016). Questions in this cluster reaffirm the interest of people with epilepsy and their loved ones in the use of medicinal marijuana for epilepsy, their concern over marijuana as a seizure trigger (≥900 users asked this type of question), and possibly a misunderstanding of the use of marijuana derivatives in treating epilepsy—many users were asked about smoking marijuana, "pot," or "weed" to control seizures, whereas it is actually being tested in a derivative, oil-based form (Epilepsy Foundation, 2016).

Cluster 13 represents users' questions related to epilepsy and sex. Sexual health issues are not well addressed in existing epilepsy self-management interventions and programs (Epilepsy Foundation, 2016). Prior studies have indicated that people, and particularly women, with epilepsy are at risk for reduced sexual quality of life (Molleken, Richter-Appelt, Stodieck, & Bengner, 2010). This cluster suggests the need to incorporate sex-related content into epilepsy self-management interventions.

Interestingly, the majority of epilepsy-related question clusters that emerged from the data set were in reference to aspects of epilepsy that are well-known issues for people with epilepsy, and information about them is readily available and included in public programs and also considered basic in patient education at epilepsy onset. For instance, the Epilepsy Foundation is an example of a highly used resource for persons with epilepsy and their families (Hesdorffer et al., 2013), and there are entire sections of the Epilepsy Foundation website, established in 2007, that contain information about common anti-epileptic drugs and their side effects, pregnancy and epilepsy, seizure triggers, definitions of seizures/epilepsy, and employment/driving issues. The Epilepsy Foundation Facebook page, established in 2009, which has more than 150,000 followers, is also continually updated with basic information about which these ChaCha users commonly asked. Both the Epilepsy Foundation website and Facebook page have been used in this manner since before 2009. There are also existing self-management interventions that address many of the issues raised by ChaCha users, though they have had limited success in affecting outcomes (DiIorio et al., 2011; Lewis, Noyes, Edwards, & Hastings, 2015). Thus, this method has allowed for a determination of gaps in knowledge in people with epilepsy that might otherwise be unknown, which is an example of how this method can contribute to furthering nursing science. Our findings also raise the question of whether Big Data can be used to identify and differentiate health-seeking behavior. Table 4 summarizes the unmet needs of this sample that we discovered in our analysis, as well as potential action items to meet those needs.

The question, then, is why and how existing resources are not meeting the needs of people with epilepsy, as well as the public. To answer this question, these initial 11 question clusters could guide deeper analysis of the content of user questions to determine the nuances of their queries that may be missed in existing interventions. It is possible that existing resources and interventions may address salient issues, but not in a relevant or detailed enough way to be beneficial, and that they may completely omit other important issues, for example, sexual health. Furthermore, these results suggest the need to review the way in which information about epilepsy is delivered to persons with epilepsy and the public via medical providers and non-profit organizations. Researchers should consider development and testing of innovative ways in which to deliver epilepsy-related information and interventions that will be more useful to the public. In addition, lack of education of people with epilepsy about their condition and its treatment from their epilepsy providers has been documented (Institute of Medicine, 2012; Miller et al., 2013; Miller, Buelow, & Bakas, 2014). This lack of adequate education could be contributory to the questions posed by this group of ChaCha users. It is also

Table 4. Summary of Unmet Needs of ChaCha Users in Relation to Epilepsy.

Needs	Potential Action/Intervention
Medication information—side effects, interactions, access	Further analyses of other social media sources to search for patient-reported side effects/ interactions of epilepsy medications Revision of current medication-related information on epilepsy.com and in other self-management programs to be aligned with social media users' concerns and questions.
General information about seizures and epilepsy—definition, causes, and disease course	Revision of current information on epilepsy.com and other self-management programs to more explicitly explain the underlying causes and disease course of epilepsy. There is a need to explicitly define "epilepsy" and other common terms such as "seizure disorder," and to differentiate them from a single seizure.
Information and services related to psychiatric comorbidities such as anxiety and depression	Revision of current information on epilepsy.com and other self-management programs to specifically address anxiety and depression and its high incidence in epilepsy; inclusion of accessible mental health resources on epilepsy.com and other epilepsy self-management programs, and possibly a public awareness campaign driven by patient questions in this data set.
Information regarding seizure triggers—definitions and how to avoid	Revision of current information on epilepsy.com and other self-management programs relating to triggers, especially those specifically mentioned in these data.
Information and resources related to issues of independence—driving and employment	Revision of current information on epilepsy.com and other self-management programs regarding driving and employment issues for people with epilepsy. It may also be necessary to provide legal resources to people with epilepsy.
Information and resources related to social interactions in the context of epilepsy (e.g., alcohol consumption)	There is a need to infuse epilepsy.com and other self-management programs with information specific to these issues, and they may best be delivered via peer-to-peer interactions. For all unmet needs—alter the ways in which existing information about these needs are delivered, to make them more consumable.

interesting to consider why ChaCha users in this data sets—many of whom likely have or know someone with epilepsy—chose to use the ChaCha service as a way in which to ask important health questions as opposed to contacting their epilepsy providers. It is also possible that the majority of questions represented in the clusters emerging from the analysis are from persons (or their family members/friends) newly diagnosed with epilepsy. Prior literature on this subset of the population is scarce, but has demonstrated the need for information knowledge about the condition during the early post-diagnosis period (Unger & Buelow, 2009).

The current study is not without limitations. It is unknown if the epilepsy-related questions that were asked in ChaCha from 2009-2012 were asked by people with epilepsy, their family or friends, curious members of the public, or those attempting to discern if they or someone known to them were having seizures. These limitations do not affect the primary purpose of the current study, which was to test the potential of the WAG modeling technique to generate and answer health-related questions of importance to nursing.

It is necessary to discuss how the modularity parameter for Louvain affected our results. The modularity level of 0.48 was achieved by choosing a resolution value of 1; this manual choice of resolution value drives the nature of partitioning to produce greater or fewer resulting groups, along with a new resulting value for modularity level, itself a quality measure of the process. Varying the resolution value above and below 1 both resulted in a lower modularity level value, so 1 appears to be optimal. There could be concern regarding potential blending of topic resulting from a misspecified modularity value, such as in Clusters 4 and 5. These two clusters are certainly similar, but respectively have several subtopics that differentiate the two, resembling a fractal onto which one could zoom into any group and further subdivide it, or zoom out and combine groups that are similar. This "lumping and splitting" of topics is a common taxonomic challenge, and best served by a combination of algorithmically driven separation into groups, followed by manual, subjective, and expert review for suitability. Raising the Louvain resolution, to in turn drive fewer resulting groups, would combine Clusters 4 and 5, but would also combine others, resulting in a shorter overall list of clusters and in turn a less diverse range of topics overall. Cluster 8 hints at the opposite threshold, where lowering the Louvain resolution, and creating a larger number of resulting clusters, would in turn allow Cluster 8 to split in two ("flashing lights," "history questions"). However, that would drive other groups to also split further, perhaps resulting in 30 or more groups instead of the 15 that we have. The balance is picking an approach that will maximize the modularity value itself, as well as result in a set of clusters that is useful by being not too many nor too few to be useful.

Results of the current study will directly inform our future work. Data analyzed in this study are dated, and were generated prior to the 2012 Institute of Medicine report on epilepsy, after which researchers and organizations responded by generating new public outreach programs. It is necessary to analyze a more updated Big Data set to determine more current concerns of persons with epilepsy. To that end, we are currently analyzing continuously updated data from Twitter, Instagram, and The Epilepsy Foundation online message boards using the WAG modeling technique. The lack of results from our analysis of the 12 million query ChaCha data set indicates that deeper analysis and modeling of the data set is likely necessary to uncover questions more loosely related to epilepsy, if they do indeed exist, and this will be one other focus of our future work. Particularly, there is a need to search specifically for epilepsy stigma-related questions in the ChaCha data set, as current analyses did not reveal any stigma-related questions, though this is a pressing issue for people with epilepsy (Epilepsy Foundation, 2016).

Acknowledgments

Access to this data set was generously made available to the Social Network Health Research Lab at the Indiana University School of Nursing by ChaCha.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Indiana University Center for Enhancing Quality of Life.

References

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *10*, 1000.

Burris, S. (2015). Public health law monitoring and evaluation in a big data future. *Journal of Law and Policy for the Information Society*, *11*, 115-125.

Burt, R. (1978). Applied network analysis: An overview. *Sociological Methods Research*, *7*, 123-130.

Carl, J. S., Weaver, S. P., & Edgerton, L. (2008). Effect of antiepileptic drugs on oral contraceptives. *American Family Physician*, *78*, 634-635.

Centers for Disease Control and Prevention. (2016). *Chronic disease prevention and health promotion*. Retrieved from http://www.cdc.gov/chronicdisease/about/prevention.htm

Chan, E. H., Sahai, V., Conrad, C., & Brownstein, J. S. (2011). Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Neglected Tropical Diseases*, *5*, e1206.

Correia, R. B., Li, L., & Rocha, L. M. (2016). Monitoring potential drug interactions and reactions via network analysis of Instagram user timelines. *Pacific Symposium on Biocomputing*, *21*, 492-503.

DiIorio, C., Bamps, Y., Walker, E. R., & Escoffery, C. (2011). Results of a research study evaluating WebEase, an online epilepsy self-management program. *Epilepsy & Behavior*, *22*, 469-474.

Epilepsy Foundation. (2016). *Epilepsy facts and statistics*. Available from http://www.epilepsy.com/

Fisher, R. S., van Emde Boas, W., Blume, W., Elger, C., Genton, P., Lee, P., & Engel, J., Jr. (2005). Epileptic seizures and epilepsy: Definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). *Epilepsia*, *46*, 470-472.

Fox, S., & Duggan, M. (2013). *Health online 2013*. Pew Internet & American Life Project. Retrieved from http://bibliobase.sermais.pt:8008/BiblioNET/Upload/PDF5/003820.pdf

Hesdorffer, D. C., Beck, V., Begley, C. E., Bishop, M. L., Cushner-Weinstein, S., Holmes, G. L., . . . Austin, J. K. (2013). Research implications of the Institute of Medicine Report, epilepsy across the spectrum: Promoting health and understanding. *Epilepsia*, *54*, 207-216.

Huang, T. T., Drewnowski, A., Kumanyika, S. K., & Glass, T. A. (2009). A systems-oriented multilevel framework for addressing obesity in the 21st century. *Preventing Chronic Disease*, *6*, A82.

Institute of Medicine. (2012). *Epilepsy across the spectrum: Promoting health and understanding*. Washington, DC: National Academy Press.

Lewis, S. A., Noyes, J., Edwards, N., & Hastings, R. P. (2015). Systematic review of epilepsy self-management interventions integrated with a synthesis of children and young people's views and experiences. *Journal of Advanced Nursing*, *71*, 478-497.

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., . . . Green, E. D. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: Capitalizing on biomedical big data. *Journal of the American Medical Informatics Association*, *21*, 957-958.

Miller, W. (2014). Patient-centered outcomes in older adults with epilepsy. *Seizure*, *23*, 592-597.

Miller, W., Bakas, T., & Buelow, J. (2013). Problems, needs, and useful strategies in older adults self-managing epilepsy: Implications for patient education and future intervention programs. *Epilepsy & Behavior*, *5*, 25-30.

Miller, W., Buelow, J., & Bakas, T. (2014). Older adults and epilepsy: Experiences with diagnosis. *Journal of Neuroscience Nursing*, *46*, 1-9.

Miller, W., Lasiter, S., Bartlett Ellis, R., & Buelow, J. (2015). Chronic disease self-management: A hybrid concept analysis. *Nursing Outlook*, *63*, 154-161.

Molleken, D., Richter-Appelt, H., Stodieck, S., & Bengner, T. (2010). Influence of personality on sexual quality of life in epilepsy. *Epileptic Disorders*, *12*, 125-132.

Murray, C. J., Atkinson, C., Bhalla, K., Birbeck, G., Burstein, R., Chou, D., . . . Wulf, S. (2013). The state of US health, 1990-2010: Burden of diseases, injuries, and risk factors. *Journal of the American Medical Association*, *310*, 591-608.

Patient-Centered Outcomes Research Institute. (2016). *Patient-centered outcomes research institute*. Available from http://www.pcori.org/

Paul, M. J., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* (ICWSM). Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2880/3264

Priest, C., Knopf, A., Groves, D., Carpenter, J. S., Furrey, C., Krishnan, A., & Wilson, J. (2016). Finding the patient's voice using Big Data: Analysis of users' health-related concerns in the ChaCha question-and-answer service (2009-2012). *Journal of Medical Internet Research*, *18*, e44.

Shaw, J. (2014, March-April). Why "Big Data" is a big deal. *Harvard Magazine*. Retrieved from http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal

Signorini, A., Serge, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the Influenza A H1N1 pandemic. *PLoS ONE*, *6*, e19467.

Smith, A. (2015). *U.S. smartphone use in 2015*. Pew Internet & American Life Project. Retrieved from http://www.pewinternet.org/files/2015/03/PI_Smartphones_0401151.pdf

Unger, W., & Buelow, J. (2009). Hybrid concept analysis of self-management in adults newly diagnosed with epilepsy. *Epilepsy & Behavior*, *14*, 89-95.

Wong, C., Harrison, C., Britt, H., & Henderson, J. (2014). Patient use of the internet for health information. *Australian Family Physician*, *43*, 875-877.